ISSN:2320-0790

COMPUSOFT
An International Journal of Advanced Computer Technology

Clustering Techniques for Streaming Dynamic Nature of Data

Swapna Vanguru[1], Anusha Merugu[2], Y.Geetha Reddy[3]

Assistant Professor, Sri Venkateswara Engineering College, Suryapet

Abstract: Nowadays many applications are generating streaming data for an example real-time surveillance, internet traffic, sensor data, health monitoring systems, communication networks, online transactions in the financial market and so on. Data Streams are temporally ordered, fast changing, massive, and potentially infinite sequence of data. Data Stream mining is a very challenging problem. This is due to the fact that data streams are of tremendous volume and flows at very high speed which makes it impossible to store and scan streaming data multiple time. Concept evolution in streaming data further magnifies the challenge of working with streaming data.

Clustering is a data stream mining task which is very useful to gain insight of data and data characteristics. Clustering is also used as a pre-processing step in over all mining process for an example clustering is used for outlier detection and for building classification model. In this paper we will focus on the challenges and necessary features of clustering techniques for streaming dynamic nature of data. Streaming data behaviour keeps on changing over time. Clustering model developed on partial data stream must be updated with new incoming data.

Keywords: Streaming data, Data Stream mining, Dynamic data, Clustering.

I. INTRODUCTION

Data streams are temporally ordered, fast changing, massive, and infinite sequence of data objects [1]. Unlike traditional data sets, it is impossible to store an entire data stream or to scan through it multiple times due to its tremendous volume. New concepts may keep evolving in data streams over time. Evolving concepts require data stream processing algorithms to continuously update their models to adapt to the changes.

Data streams are ubiquitous. These can be found in many application domains from online financial transaction to medical systems and space research centers, where satellites are continuously generating streaming data. And even new applications are emerging day by day due to significant growth in computer processing speed and spread of computer networks. So there is a need of effective and efficient data mining techniques for streaming data which can handle the challenges associated with streaming data. Data mining techniques for streaming data includes: clustering, classification, frequent pattern mining and outlier detection which can be used to mine patterns from streaming data. In this paper, we are focusing on clustering. Because it is helpful to gain insight of data and data characteristics and can be used as a pre-processing step with other data mining techniques.

II. DATA STREAM CLUSTERING

Data stream clustering discovers clusters in the streaming data. Data stream clustering is very different from traditional clustering: For traditional clustering data sets are static, but the data stream is dynamic in nature. Because of massive size of data stream, it is not possible to store data streams in memory and scan it multiple times. But for traditional clustering data sets are store in memory and can be scanned multiple times. The clustering results of data streams change over time, but not so for traditional clustering. There are following two different approaches for data stream clustering: Data stream clustering by example and data stream clustering by variable.

III. TECHNIQUES FOR DATA STREAM CLUSTERING BY EXAMPLE

Data stream clustering by example treats data points coming at same time stamp from different data source as one unit in clustering process. Each data unit describes the features of an entity at a particular time stamp. For an example in a bank daily, there are many transaction that are on different account numbers, are of different type like credit or debit by cash, check or online transaction and also other features are associated like avail balance, amount deposited or withdrawn etc. so for a single transaction in this case many data points are recoded that are of different types and in combination form a data unit.

A. Techniques For Data Streams Clustering By Example

STREAM is an algorithm for data streams clustering. It consists of two phases and follows divide and conquer approach. In first phase, it divides the data streams in buckets and then finds k clusters in each bucket by applying k-median clustering. It stores cluster centres only and cluster centres are weighted based on the number of data points belongs to corresponding cluster and then discard the data points. In second phase weighted cluster centres are clustered in small number of clusters. Though its space and time complexity is low but it cannot adapt to concept evolution in data. CluStream [4] is proposed by Aggarwal et al. It divides the clustering process in following two online component and offline components. Online component stores the summary of data in the form of micro-clusters. Micro-cluster is the temporal extension of clustering feature of BIRCH [5]. Summary statistics of data are stored in snapshots form which provides the user flexibility to specify the time interval for clustering of micro-clusters. Offline component apply the kmeans clustering to cluster micro-clusters into bigger clusters.

D-stream [8] is a density based grid clustering algorithm for streaming data. It divides the complete data space in grids. It also comprises two phases. In online phase, it maps incoming data point on the corresponding grid. In offline phase it calculates density of each grid and then discards the data. For final clusters it clusters the grids based on their density. It uses fading function to decrease the density of grids with time, if it falls below a threshold and no new data point is added since last checking of grid density that grid is discarded. But it is not scalable on number of data dimensions because with increase in number of dimension number of grids increase exponentially.

E-Stream [10] is a data stream clustering technique which supports following five type of evolution in streaming data: Appearance of new cluster, Disappearance of an old cluster, Split of a large cluster, merging of two similar clusters and change in the behaviour of cluster itself. It uses a fading cluster structure with histogram to approximate the streaming data. Though its performance is better than HPStream algorithm but it requires many parameters to be specified by user.

IV. Techniques for Data Stream Clustering By Variable

Data stream clustering by variable treats data points coming from a single source as a stream of data and all the data points of same data stream must belong to the same cluster. For an example in medical domain, patient health monitoring system continuously generates various streams of data which continuously records heart bit rate, blood pressure, body temperature and other clinical measurements of patient.

B. Applications for Clustering Streaming Data

- Application in Stock Exchange – Another application of data stream clustering is in stock exchange. Price of stocks keeps on rise and fall with time. But some stocks price rise and fall concurrently in some time intervals. Such stocks can be grouped together using stream clustering technique. This information will be useful for investors.
- Application in Meteorological Research - Weather forecasters cluster the data streams that are collected from various geographical locations to identify regions with similar characteristics.
- Application in Supermarket – Various supermarkets record their daily sales of various items. Similarly behaving items can be clustered together by analysing their sales relation and with this information, supermarket can maximise the profit by manipulating the price of items like by discounts on product combination.

C. Tools for Clustering Streaming Data

Massive Online Analysis (MOA) [21] - It is software which provides the support for implementation of algorithms and to conduct experiments for clustering and classification of evolving data streams. It contains a collection of various algorithms for both classification and clustering as well as related datasets and measure for evaluation of clustering and classification model. MOA and related material can be downloaded from following link: http://moa.cms.waikato.ac.nz/downloads/

V. References

[1] J. Han and M. Kamber, Data Mining: Concepts and Techniques, J. Kacprzyk and L. C. Jain, Eds. Morgan Kaufmann, 2006, vol. 54, no. Second Edition.

[2] Yogita and D. Toshniwal, “A framework for outlier detection in evolving data streams by weighting attributes in clustering,” in Proceedings of the 2nd International Conference on Communication Computing and Security, India, 2012.

[3] L. callaghan, N. Mishra, A. Meyerson, S. Guha, and R. Motwani, “Streaming-Data Algorithms for High-Quality Clustering,” in Proceedings of IEEE International Conference on Data Engineering, 2001.

[4] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, “A framework for clustering evolving data streams,” in Proceedings of the 29th international conference on Very large data bases - Volume 29, ser. VLDB ’03. VLDB Endowment, 2003, pp. 81–92.

[5] T. Zhang, R. Ramakrishnan, and M. Livny, “Birch: an efficient data clustering method for very large databases,” in Proceedings of the 1996 ACM SIGMOD international conference on Management of data, ser. SIGMOD ’96, New York, NY, USA, 1996, pp. 103–114.

[6] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, “A framework for projected clustering of high dimensional data streams,” in Proceedings of the Thirtieth international conference on Very large data bases – Volume 30, ser. VLDB ’04. VLDB Endowment, 2004, pp. 852–863.

[7] F. Cao, M. Ester, W. Qian, and A. Zhou, “Density-based clustering over an evolving data stream with noise,” in SIAM International Conference on Data Mining, 2006.

[8] L. Li-xiong, H. Hai, G. Yun-fei, and C. Fu-cai, “rdenstream, a clustering algorithm over an evolving data stream,” in International

Conference on Information Engineering and Computer Science, 2009, 2009, pp. 1–4.

[9] Y. Chen and L. Tu, “Density-based clustering for real-time stream data,” in Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, ser. KDD ’07, New York, NY, USA, 2007, pp. 133–142.